Capital Bike Share Predictive Model Report

Prepared By: Clark P. Necciai Jr.

November 18, 2023

Contents

1 Executive Summary

Capital Bike Share provides a network of bicycles to the denizens of the Washington D.C. region. We were approached by Capital Bike Share to delve into their dataset observations across 2011 and 2012. Preliminary insights revealed that rental demand can be affected by a variety of influential factors.

We examined multi-variable trends and deliberated on insightful patterns. Based on our early findings, we suspected rentals being driven by bicycles being used as a primary mode of transportation during work-commute hours. Furthermore, our resulting model confirmed our preliminary analysis, as it was ultimately determined that hour variables and temperature were among the most significant factors in accurately predicting bicycle rentals.

2 Problem Statement and Approach

Our primary tasks in this analysis were twofold:

- Identifying the most influential predictive variables in determining total bicycle rentals
- Fitting a multiple regression model predicting total bicycle rentals

Beginning with an exploratory data analysis, we determined variables which had significant relation to the target variable. After a well-documented, granular analysis of 17 features over 17,379 observations, we utilized our findings for comprehensive modeling utilizing stepwise selection to select only the most significant predictors. This process culminated in a multiple linear regression model that not only accurately predicted rental demand, but simultaneously provided assumption-backed diagnostics confirming our models' reliability and generalizability. We then end with our conclusions and recommendations.

3 Methodology

3.1 Data Preprocessing

We began our approach with an overview of the integrity of our dataset's structure. We discovered no missing values nor duplicated observations[Appendix-Fig.1]. We did, however, note variables which had data types that were non-representative of their underlying values.

Variables `dteday`, `season`, `yr`, `mnth`, `hr`, `holiday`, `weekday`, `workingday`, and `weathersit` were all considered to have data types and values which were non-representative. We applied more representative data types and applied appropriate labels[Appendix-Fig.2/3]. `dteday`, `yr`, `mnth`, `hr`, `weekday`, `weathersit`, and `cnt` were renamed for additional clarity.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Target Variable

3.2.1.1 `count_rentals` The variable we aim predict is the total hour-by-hour number of bicycle rentals, `count_rentals`. Our target variable's distribution shows us that

`count_rentals` is highly right skewed[Appendix-Fig.4]. There is serious variability in the hour-by-hour rentals, with a minimum of a single rental to a maximum of nearly a thousand rentals per hour, but with 50% of our observations being less than the median of **142** rentals.

3.2.2 Predictor Variables

Each categorical, continuous, and Boolean features' summary statistics and distributions were investigated. Continuous features considered were `temp`, `atemp`, `hum`, and `windspeed`[Appendix-Fig.5]. Boolean features were `holiday` and `workingday`[Appendix-Fig.6]. Categorical features were `season`, `year`, `month`, `day`, `hour`, `date`, and `weather`[Appendix-Fig.7/8].

3.2.2.1 Noteworthy Discoveries

3.2.2.1.1 `date` The `date` distribution revealed that not every unique value of date has an expected equivalent number of hourly interval measurements. The majority of each of the **731** distinct date values contained twenty-four or twenty-three of the expected hourly observations. However, fourteen of the dates contained fewer, which in turn ultimately equated to **103** hours worth of missing possible observations[Appendix-Fig.9].

3.2.2.1.2 `holiday` Inspection of `holiday` revealed dates which were incorrectly labeled. We re-labeled these observations based on official federally recognized holidays.

2011-01-01, 2011-12-25, 2012-01-01, and 2012-11-11 were incorrectly mislabeled as *not* being holidays. 2011-12-26, 2012-01-02, and 2012-11-12 were mislabeled as **being** holidays. 2011-04-15 and 2012-04-16 are dates of observance for Emancipation Day in the Washington D.C. Area. Due to the holiday-like observance of Emancipation Day in our area of interest, these two dates will be labeled as holidays.

3.3 Correlation

We investigated the extent to which our variables are linearly related to one another with a correlation matrix[Appendix-Fig.10]. Here, we are concerned with relationships between that of our target, `count_rentals` and continuous predictors. Our noteworthy findings include:

- `temp` and `atemp` both aim to measure normalized Celsius in an objective and subjective way, respectively. It is unsurprising that these variables exhibit a near exact linear relationship with one another. We see they are both equally, positively correlated to our target variable. Naively including both `temp` and `atemp` in our modeling process would result in multicollinearity. To maintain a robust regression model and avoid multicollinearity which would cause our coefficients estimates to become unstable, we will choose only `temp` to be included in the modeling process due to it being an objective measurement as opposed to `atemp` which is subjective.
- Our target is negatively correlated with `hum`. Being that humidity includes precipitation such as rain, mist, snow, and others, this relationship is unsurprising.

Overall, we see that each of our continuous predictor variables exhibits a mostly straight-line relationship with our target variable[Appendix-Fig.10 (Bottom Row)].

3.4 Multivariate Analysis

3.4.1 Temperature and Target Relationship

We examined the multi-variable relationship between our temperature related variables, `temp`, `atemp`, and `count_rentals`[Appendix-Fig.11]. We revealed twenty-four high-leverage observations having the exact same `atemp` value of 0.2424, but `temp` values which were relatively high and variable in relation, ranging from [0.62 to 0.86]. When considering the observational evidence of these twenty-four `temp` and `atemp` observations as being high leverage outliers, we have opted to remove these observations. This removal helps to ensure data integrity as we move towards the modeling process.

3.4.2 Work Commute - Hourly Rental Trends

Our intuition led us to investigate possible patterns of usage based solely on time. What we uncovered, showed that during working days, there is strong indication that bicycles are being utilized as a primary mode of transportation to and from work by `registered` riders. Around 8:00AM, 5:00PM, and 6:00PM, we note high rental demand during the busiest work commute times, with a stark decrease afterwards. Neither `casual` rentals nor rentals occurring on Saturday/Sunday exhibit this trend, adding further evidence to our speculation[Appendix-Fig.12/13]. We see that between total rentals and registered rentals, the trend is nearly identical[Appendix-Fig.14/15].

3.5 Feature Reduction

`date` - Our team reasoned that creation of **730** dummy variables representing `date` could have severe impacts (high VIF and overfitting) on model performance. No unique information is provided by `date` that is not already found in our `year`, `month`, and `day` variable. Because of this, we have decided to disregard the `date` variable from the modeling process.

Additionally, being that `instant` is merely an identifier, it will also be disregarded. Lastly, being that our target variable is the exact sum of `casual` and `registered`, inclusion of these two would make our predictions arbitrary. We likewise disregarded these from our modeling and assume this information is unknown during predictions.

3.6 Feature Engineering

Dummy Variable Creation - Categorical variables, such as `season`, `year`, `month`, `day`, `hour`, and `weather`, are suited for dummy variable creation for use in our multiple linear regression model. This encoding ensures these variables are appropriately interpreted by our model.

4 Model Building

4.1 Data Partition

With our goals centered on finding the most significant predictors with respect to determining the total number of rentals, we decided to partition our data 80% towards training the model and 20% for testing model performance. Allowing a significant majority of our overall dataset to go towards the model fitting process enables us to estimate precise, stable coefficient estimates and better overall model performance.

4.2 Initial Full Model(s)

Our initial, comprehensive full model encompasses all possible predictors within the Capital Bike Share dataset. Our approach allows us to gain a generalized sense as to the importance and effects our predictors are having in determining our response variable.

Initial Full Model

$$\hat{count}\hat{rentals} = \hat{\beta}_0 + \hat{\beta}_1\texttt{holiday} + \hat{\beta}_2\texttt{workingday} + \hat{\beta}_3\texttt{temp} + \ldots + \hat{\beta_{52}}\texttt{weather\_Type\_4}$$

Where $\hat{count}\hat{rentals}$ is the predicted number of rentals, $\hat{\beta}_0$ is the estimated intercept, and $\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta_{52}}$ are the estimated coefficients of the predictor variables[Appendix-Pages:20/21].

While the full model containing all possible predictors can provide some insightful information, it is unrefined. Many of the predictors were found to be statistically insignificant in predicting the response. While we could remove these variables, we will allow stepwise selection to determine the best set of predictors by systematically adding and removing variables to determine the best subset of significant predictors. To refine this full model, we will first inspect the assumptions of multiple linear regression.

4.2.1 Model Diagnostics

4.2.1.1 Normality of Residuals Our residuals should be normally distributed and can be visualized using a Q-Q plot. Deviations from our straight line in the diagnostic plot would suggest potential non-normality and is confirmed by the formal Kolmogorov-Smirnov Test. We can be certain that our residuals are not normally distributed[Appendix-Page:22/Fig.16].

4.2.1.2 Homoscedasticity The homoscedasticity assumption states that we should have residuals with a constant variance. The funnel-shape we see in our diagnostic plot is indicative of the opposite, heteroscedasticity, or non-constant variance, and is confirmed by our formal Breusch-Pagan Test. Our standard errors, confidence intervals, and subsequently our hypothesis testings rely on the homoscedasticity assumption[Appendix-Page:23/Fig.17].

4.2.1.3 Linearity The linearity assumption states that we should assume the true relationship between our predictors and response variable is a straight-line. We can identify non-linear trends with the red line fit to our residuals. Linearity appears to be violated here, as the upward-curved line is indicative of a non-linear relationship[Appendix-Fig.17].

4.2.1.4 Multicollinearity The presense of multicollinearity reduces the accuracy in our model's coefficients by causing our coefficients' standard errors to grow, effectively masking their importance and making interpretation difficult. We can detect multicollinearity by calculating the variance inflation factor of our model's predictors. Our first full model contains a few predictors having variance inflation values greater than 10 which warrants possible removal from our model to address multicollinearity[Appendix-Page:24/25].

4.2.1.5 Independence of Errors Our residuals should be independent of each other for our regression to be reliable. We can see from the Residuals vs. Fitted plot's red line that we have evidence of having violated this assumption. We can confirm that indeed we have a lack independence of errors from the formal Durbin-Watson test. From our diagnostic plot, if our red line were horizontal, this would indicate independence of errors. Our test result shows that indeed there is strong evidence of positive autocorrelation and a lack of independence of errors[Appendix-Fig.17].

4.2.2 Model Improvement

We want our final multiple linear regression model to be robust. To achieve this and bring our assumptions closer to be satisfied, we now proceed with various approaches of addressing issues which may be affecting our model's assumptions.

4.2.2.1 Influential Observation Removal We begin by removing observations which we believe may be having too disproportionate an impact on our model's fit. These observations in particular are likely to disproportionately affect our model's coefficients, affecting our overall predictive accuracy. These include the three observations with significantly high Cook's distances which measures an observation's leverage. Each removed observation were of `weather_type_4`. Because of this, the `weather_type_4` variable was consequently removed from the training data set. Being that there were a mere three observations of `weather_type_4` observed over the entire dataset, we feel justified in removal of this variable and its disproportionate impact on our model outside of normal weather conditions[Appendix-Fig.18].

4.2.2.2 Target Variable Transformation We considered multiple transformations to approach normality and address the linearity and homoscedasticity assumptions, including, the Box-cox, logarithmic, square-root, cube-root, and fourth-root transformations. The transformation which approximated normality (skewness value to 0) the closest was the **cube-root** transformation. Therefore, we will predict the cube-root transformed version of our target variable `cube_root_total` for the final model[Appendix-Fig.19].

4.2.2.3 Multicollinearity We opted to remove the influences of multicollinearity in our models by removing the variables which had the highest variance inflation factors(VIF > 10) one at a time. The variables which were removed were `workingday`, and `season_Summer`.

4.3 Final Reduced Model Using Stepwise Selection

Using stepwise selection, we identified an optimal subset of predictors. These variables were determined to be the best set of predictors in predicting the transformed response, `cube_root_total`[Appendix-Page:28/29]. The final stepwise model formula can be found in the Appendix on Page 30.

Our team believes that two of the best metrics for determining model fit are Adjusted R^2 and Bayesian Information Criterion (BIC) among others. Both of these metrics help determine models which balance goodness-of-fit and complexity. We seek to maximize Adjusted R^2 and minimize BIC through stepwise selection's repeated addition and removal of predictors to determine best fit[Appendix-Fig.21].

4.3.1 Interpretation of Model Coefficients

From our step-wise model, the three variables with the most significant impact on the expected number of total rentals are `hour_5:00PM`, `hour_6:00PM`, and `hour_8:00AM`.

For instance, when our `hour_5:00PM` variable is 1(True), holding other variables constant, the cube root number of bicycle rentals is expected to increase by 4.23614. In other words, if we back-transform our coefficient values for interpretability, the number of bicycle rentals is expected to increase by approximately 77 total bicycle rentals when it is 5:00pm. This procedure of coefficient interpretation through back-transformation applies to all model coefficients. In the case of continuous variables, such as `temp`, the expected increase in the number of bicycle rentals is determined by a one unit increase in the predictor value.

4.3.2 Model Diagnostics

Our assumptions mentioned after the first full model fitting were promptly addressed with our transformation of the response variable via a cube root transformation, and removal of over-influential observations and high variance inflation variables. Despite our approach, we did observe some potential violations of regression assumptions in homoscedasticity, linearity, and normality/independence of residuals. The multicollinearity assumption was satisfied due to our removing of high variance inflation variables[Appendix-Pages:31/32].

In practice, however, the assumptions are rarely validated. After employing our corrective measures and observing the vastly improved diagnostic plots, we remain confident in our model's predictive capabilities and generalizability. Our attempts at rectifying our assumptions helped to stabilize our coefficients, and ensure that our model contains meaningful variables in determining our target[Appendix-Fig.22].

4.3.3 Test Set Evaluation - Model Performance

Utilizing the final step-wise model determined above, we tested our model's performance by making predictions on the test set. Because we applied a cube-root transformation to our target variable `cube_root_total`, we back-transform these predicted values for interpretability. Doing so allows us to draw meaningful interpretation from our model's performance in the original units as the response.

Here we discuss the test performance metrics of Root Mean Squared Error(RMSE), R^2, and Mean Absolute Error(MAE). RMSE and MAE are measured in the same units as the original target variable and thus have a meaningful interpretation. MAE for example, is the "on-average" error between the predicted value and the true number of bike rentals, which for our model is approximately **63** rentals. Additionally, our R^2 value is approximately 0.733, meaning that 73.3% of the variability in the number of bike rentals is explained by our model. Our RMSE is approximately 95.5, which is indicative of strong, predictive accuracy in our model.

From our visualization displaying predictions versus true observational values, we see that our residuals are fairly small when the number of actual rentals are low. The residuals slightly increase in magnitude as we make predictions for larger numbers of rentals which inflates our performance metrics[Appendix-Fig.23]. Overall, we believe this model to be of high quality and generalizable to new data.

5 Conclusions

The Capital Bike Share dataset analysis revealed insightful patterns and statistics in pursuit of variables which were used in formulating a reliable multiple linear regression model. Significant predictors in determining the total number of bicycle rentals found when examining our data, such as the hour of the day and temperature, help us understand which features highly influence the number of total rentals.

Our final model provides a concise selection of key predictors, allowing us to filter out non-meaningful features which would otherwise prove irrelevant. With our final multiple linear regression consisting of only the most paramount variables in determining the number of rentals, Capital Bike Share can feel confident in making accurate and reliable predictions.

5.1 Recommendations

Based on the hourly working day commute trends showing commutes to and from the workplace as being a major driving force in the number of rentals, and the extreme prevalence and significance of the hour variables present in our final model, we will lastly provide actionable insight for Capital Bike Share.

We are certain that registered commuters are utilizing bicycles as their primary mode of transportation during workplace commutes. Capital Bike Share can act on this, by installing/placing additional bicycle kiosks in key residential areas and near workplaces with traffic-heavy commute roads. Doing so would incentivize additional bike rentals by providing commuters a reliable source of workplace transportation, ultimately increasing the number of bicycle rentals. Additionally, to improve future modeling, we recommend gathering more observations with larger numbers of rentals, as this data would improve predictive accuracy at those larger values of rentals. A lack of high-rental observations is more than likely the cause of our large residuals when predicting high numbers of rentals.

6 Appendix

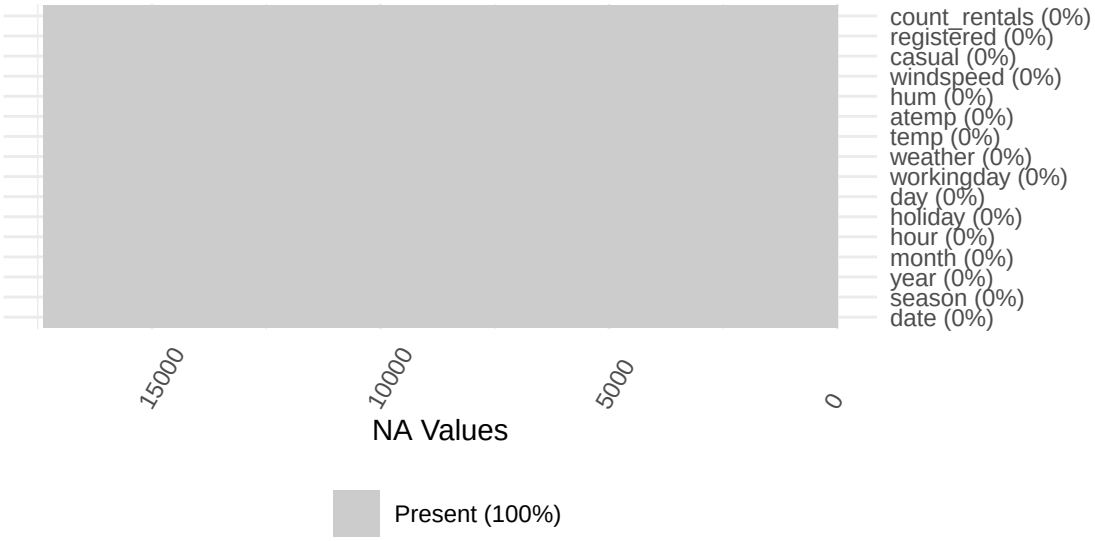

Figure 1: No missing values in dataset

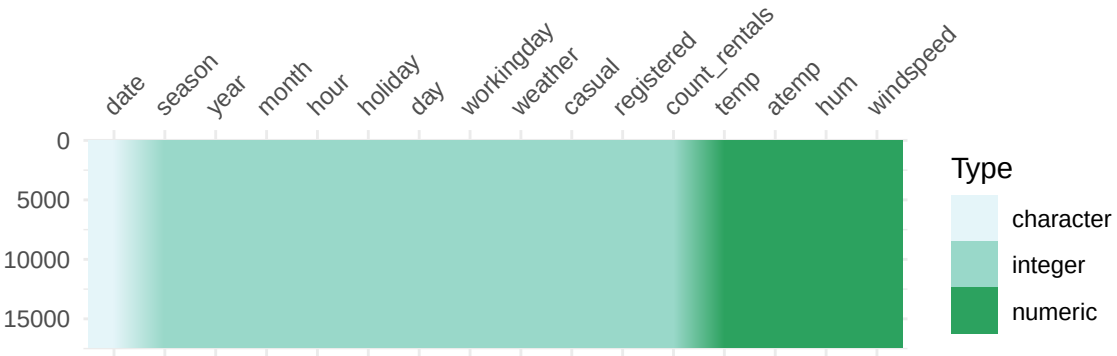

Figure 2: Pre-assigned Variable Data Types

\begin{figure}

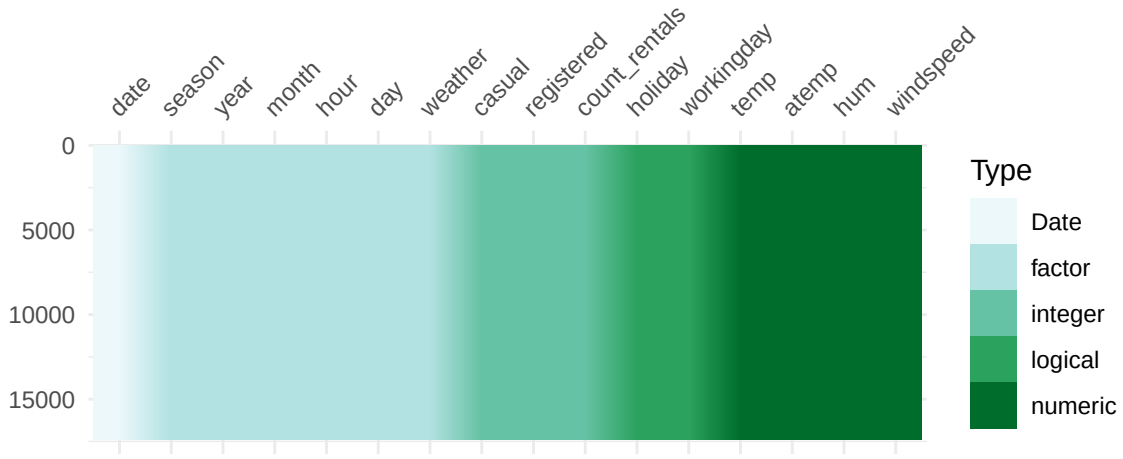

Figure 3: Reassigned Variable Data Types

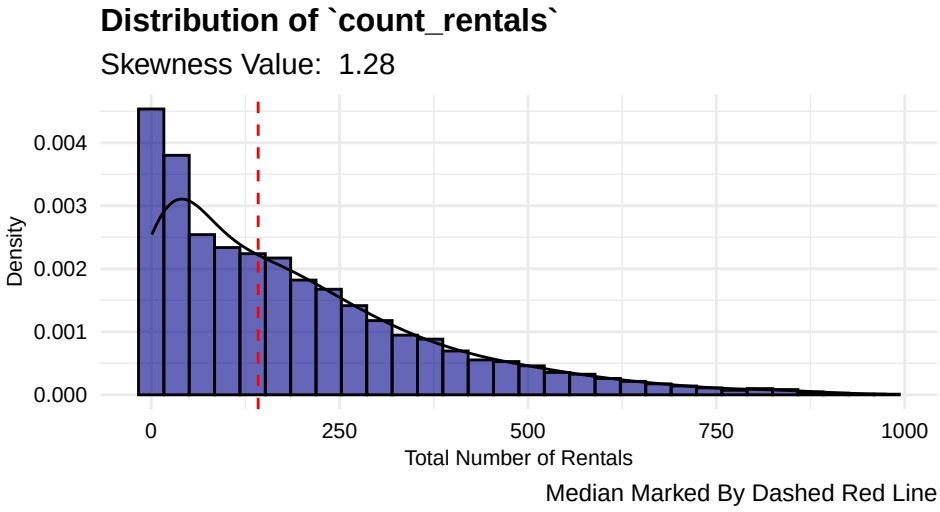

Figure 4: Distribution of Target Variable

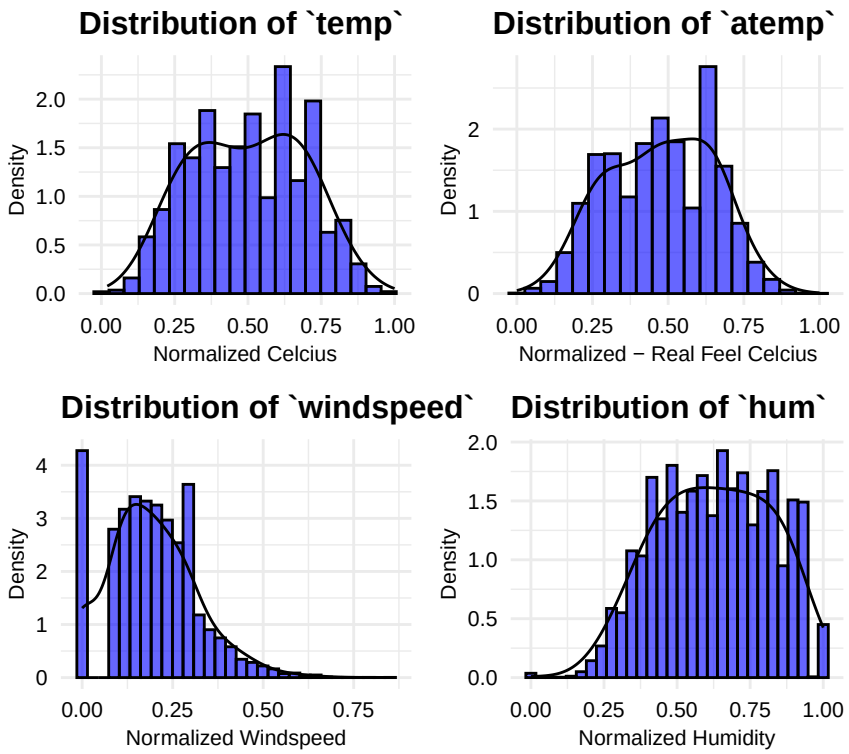

Figure 5: Continuous Variables Distribution

	Mean	Median	StdErr	Skew	Q1	Q3	IQR	Min	Max
count_rentals	189.46	142	181.39	1.28	40	281	241	1	977
temp	0.5	0.5	0.19	-0.01	0.34	0.66	0.32	0.02	1
atemp	0.48	0.4848	0.17	-0.09	0.3333	0.6212	0.2879	0	1
hum	0.63	0.63	0.19	-0.11	0.48	0.78	0.3	0	1
windspeed	0.19	0.194	0.12	0.57	0.1045	0.2537	0.1492	0	0.8507

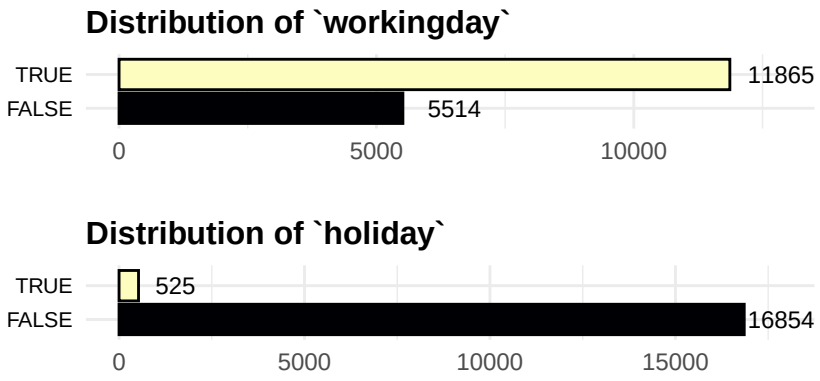

Figure 6: Boolean Variables Distribution

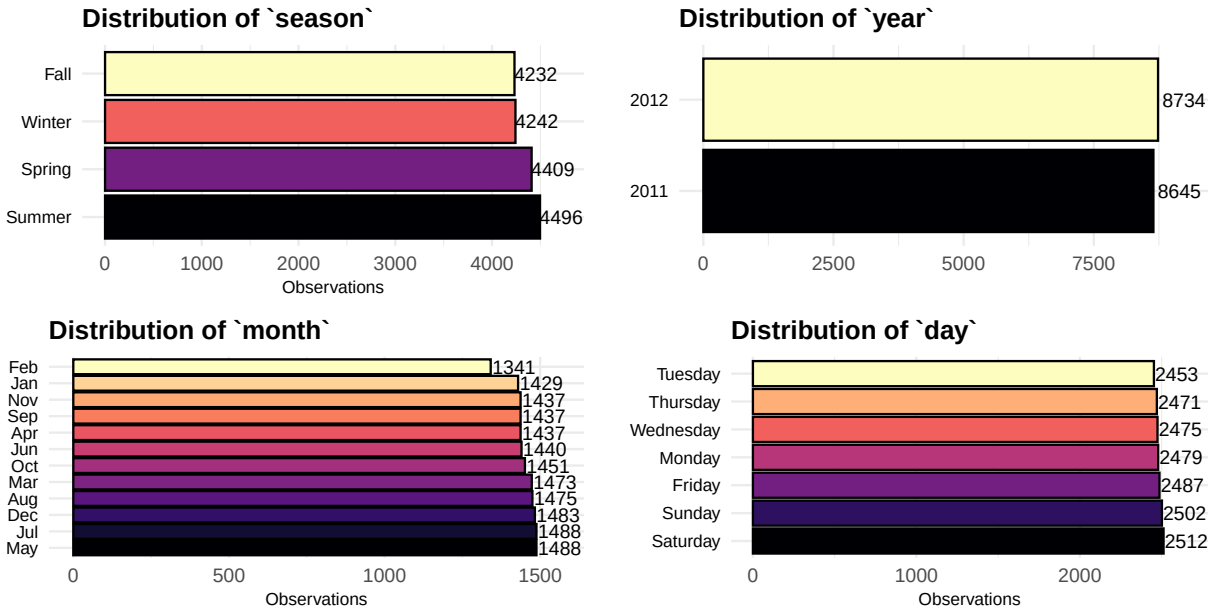

Figure 7: Distributions of ‘season‘, ‘year‘, ‘month‘, and ‘day‘

```
## NULL
```

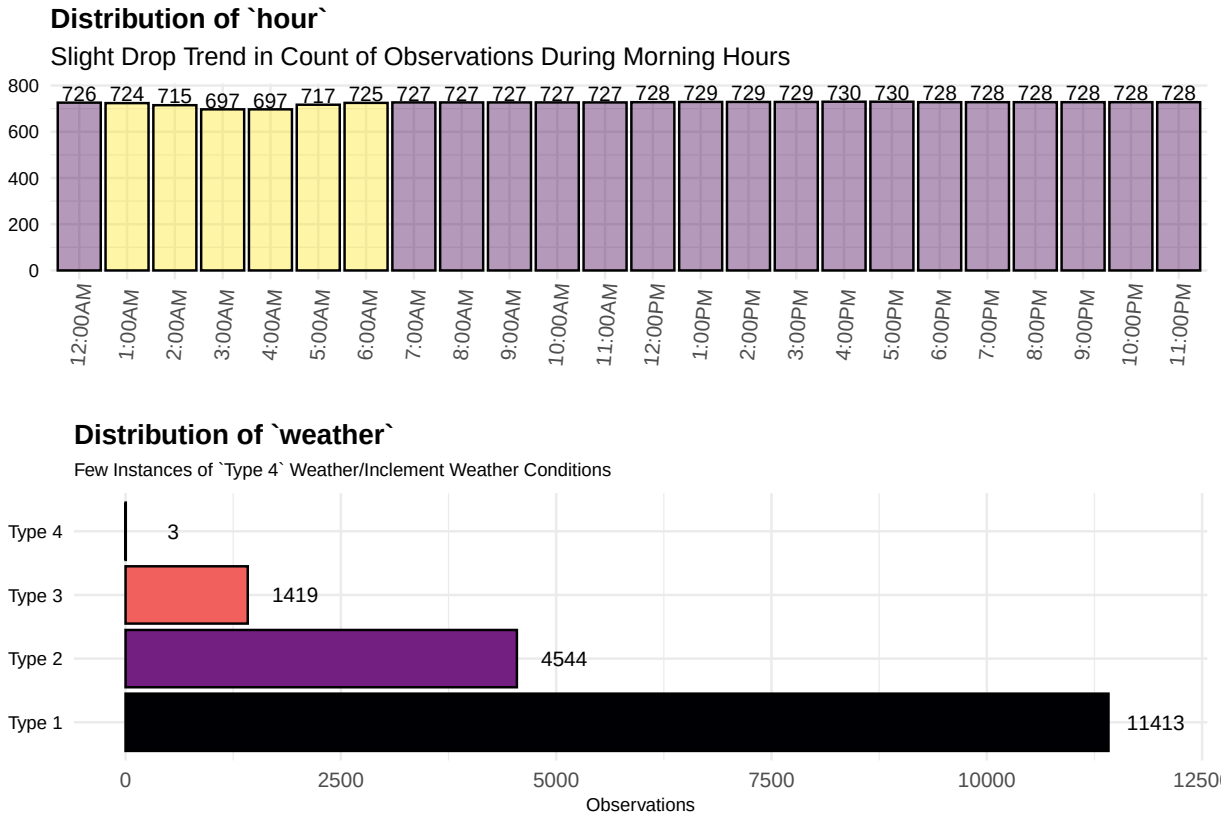

Figure 8: Distributions of 'hour', and 'weather'

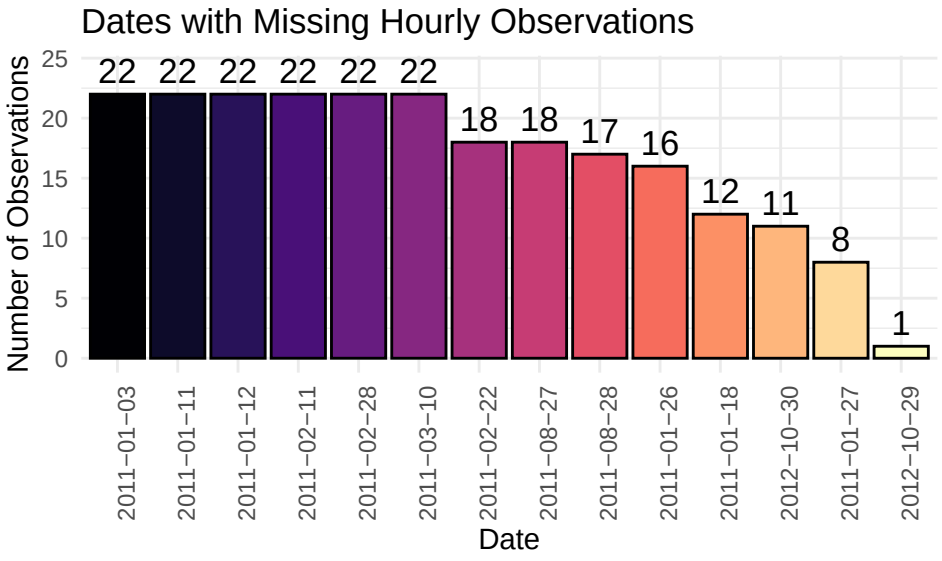

Figure 9: Dates with Low Number of Observations

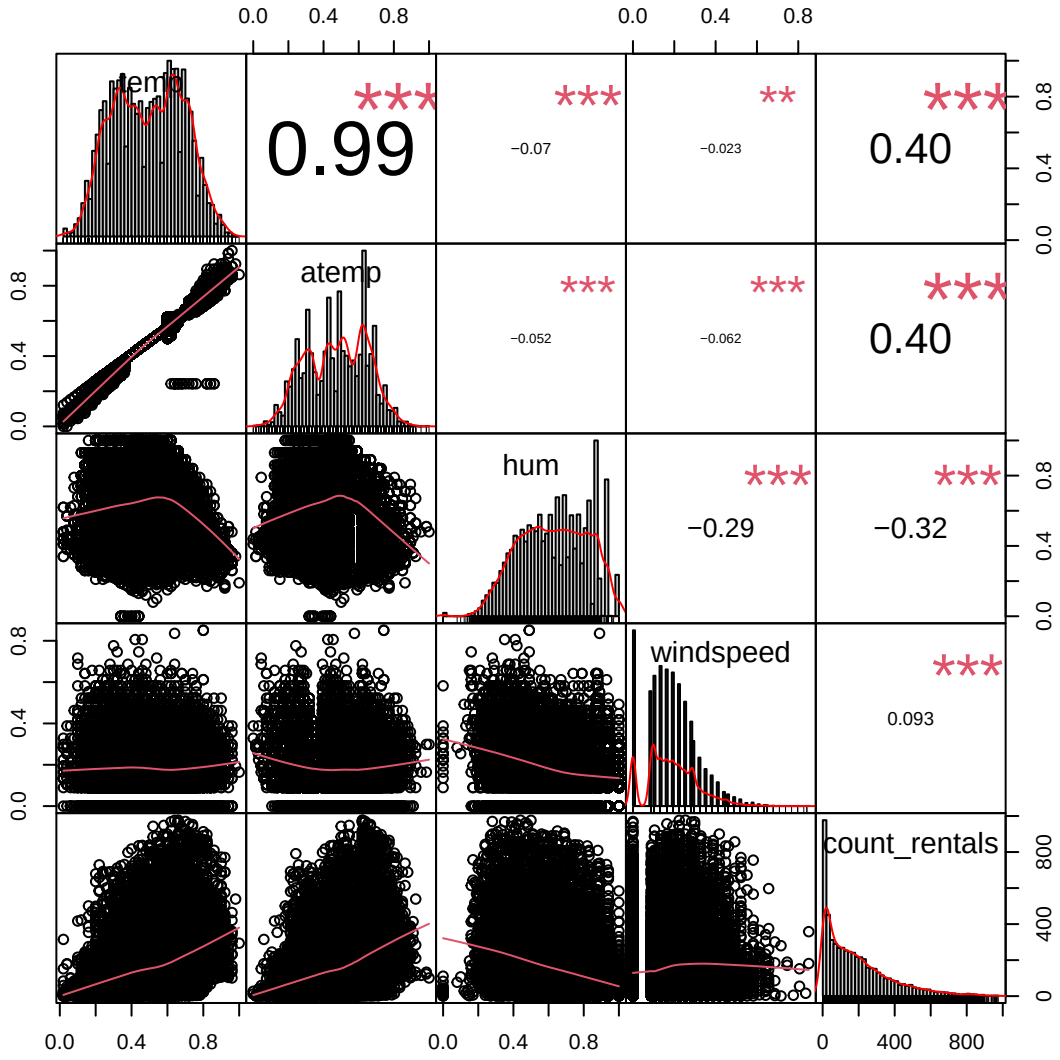

Figure 10: Correlation Matrix and Continuous Variables Relationship with Target Variable

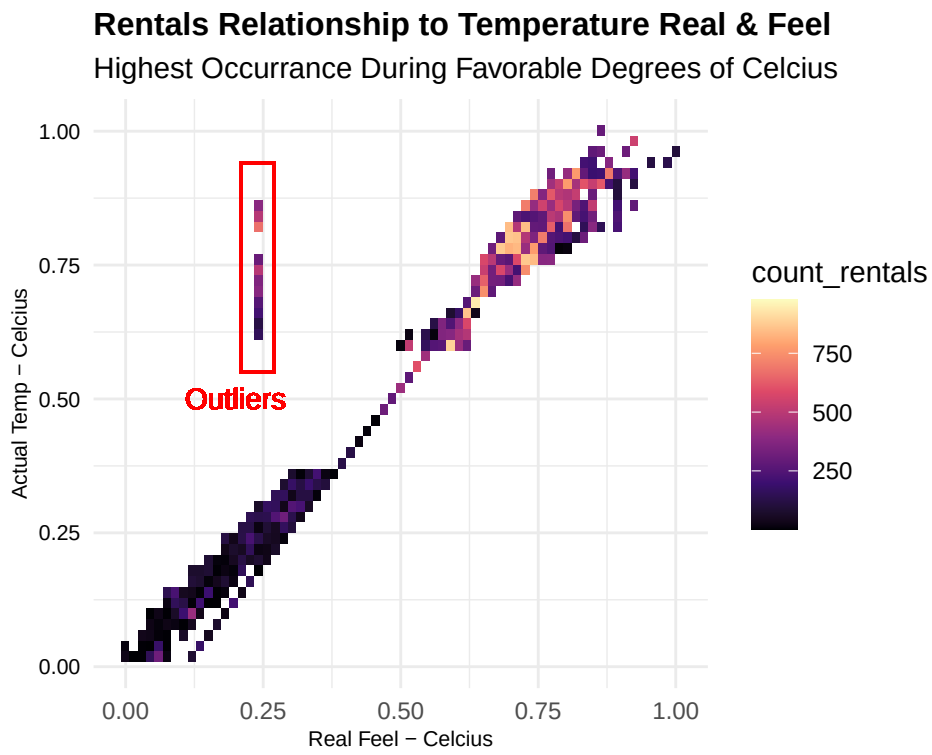

Figure 11: High Leverage Points Between ‘temp‘ and ‘atemp‘

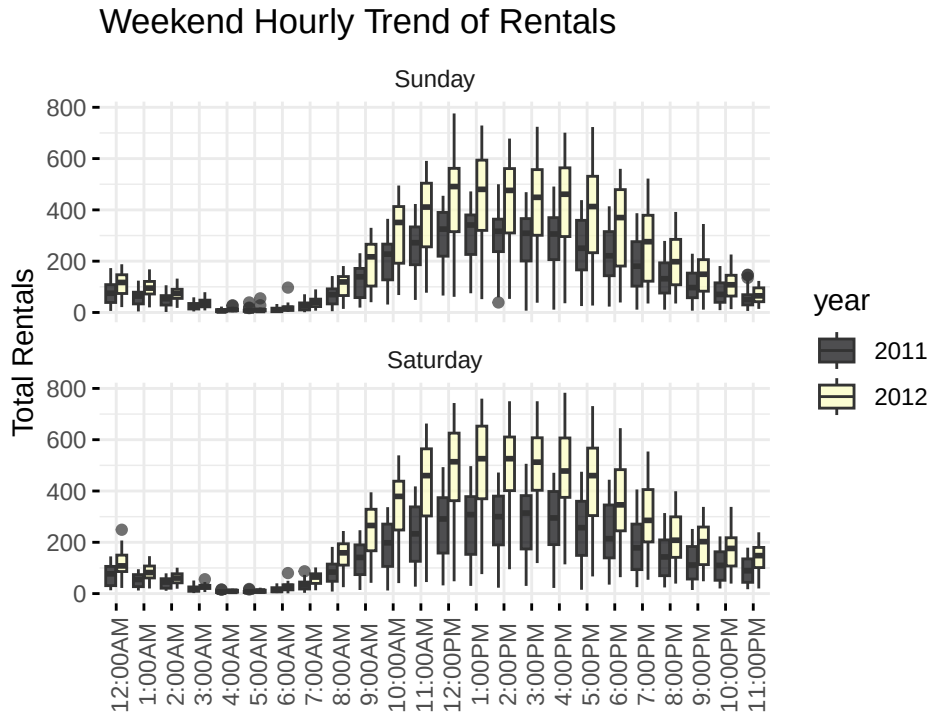

Figure 12: Smooth Curve of Rental with Peak Usage Around Midday

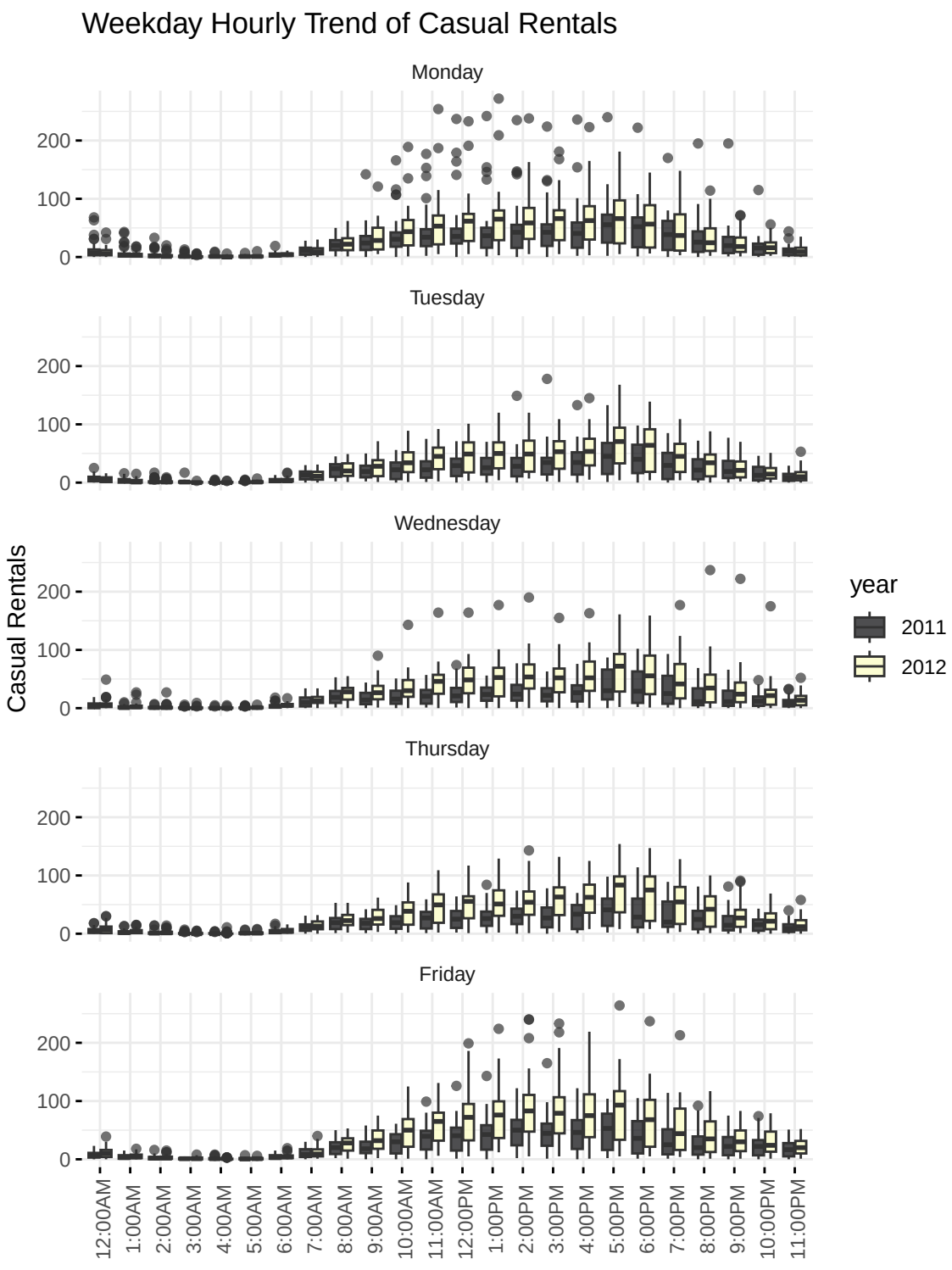

Figure 13: Low Casual Usage Throughout Working Days

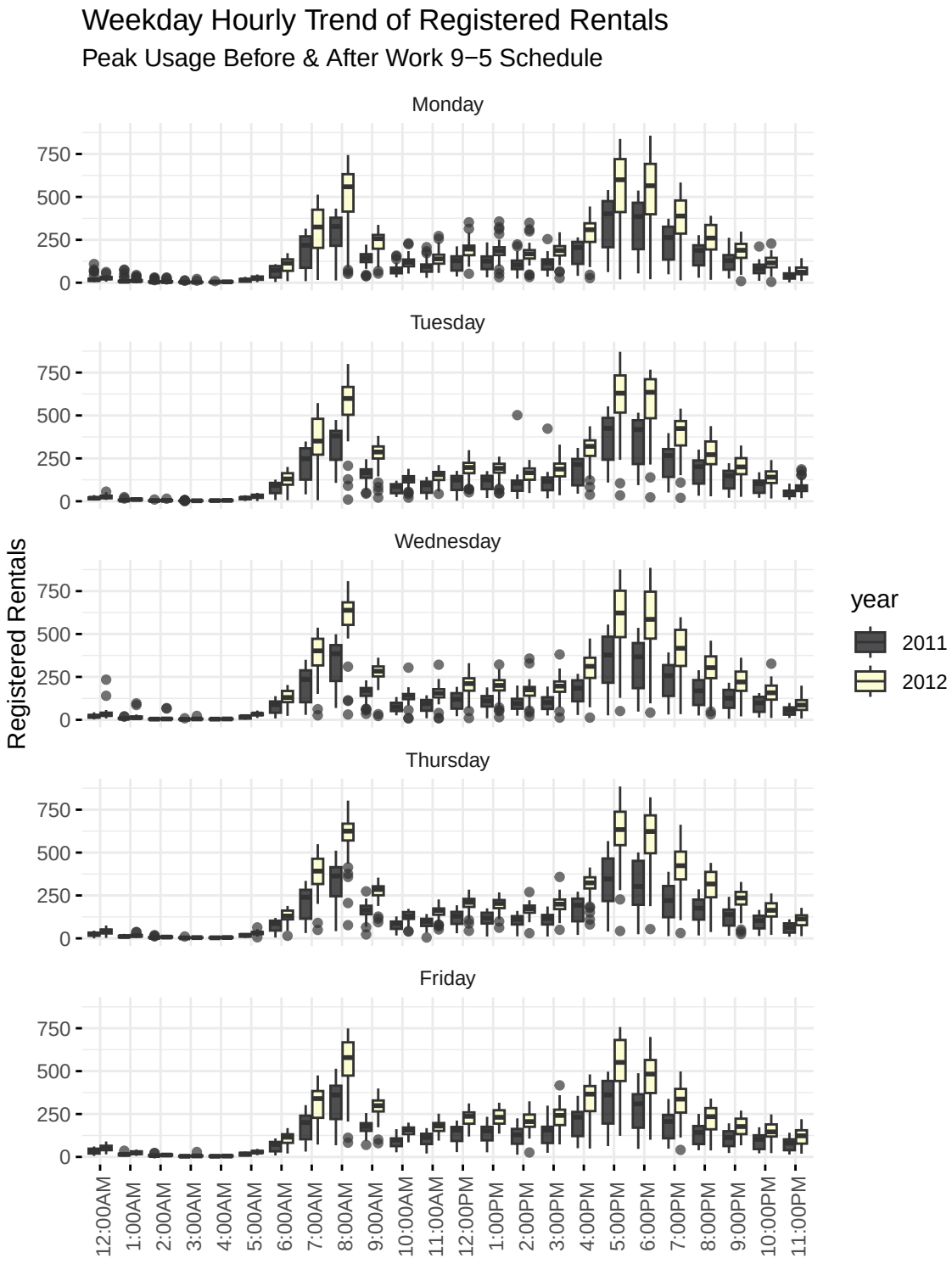

Figure 14: Registered Work Commute Usage

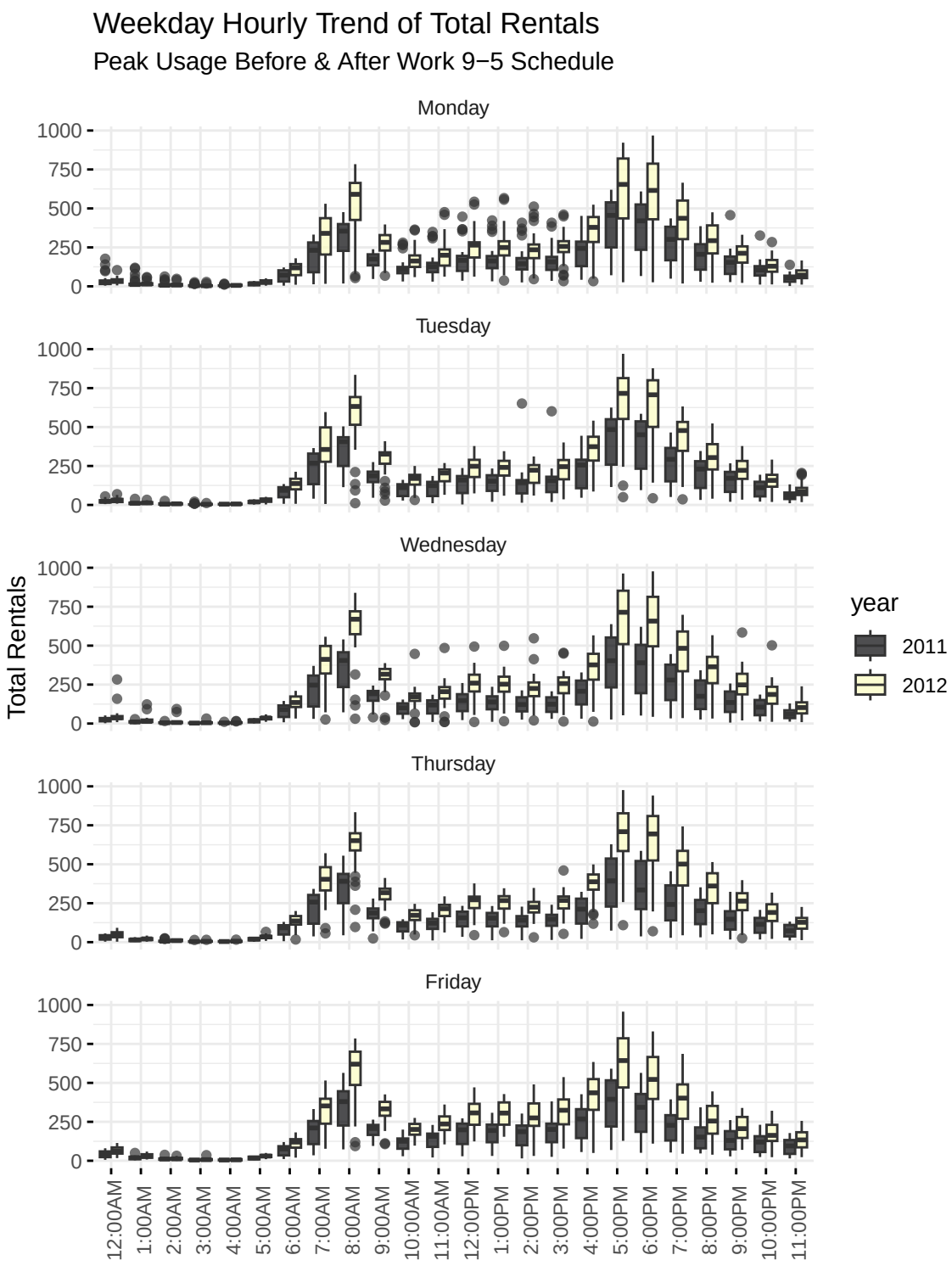

Figure 15: Total Rentals Work Commute Usage

Initial Full Model

```
##
## Call:
## lm(formula = count_rentals ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.60  -60.61   -7.50   50.84  447.87
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -97.318      7.392 -13.166  < 2e-16 ***
## holiday             2.790      9.170   0.304 0.760899
## workingday         30.070      9.596   3.134 0.001730 **
## temp              240.480     10.506  22.889  < 2e-16 ***
## hum               -80.998      6.211 -13.040  < 2e-16 ***
## windspeed         -32.335      7.682  -4.209 2.58e-05 ***
## season_Spring      37.464      5.434   6.894 5.65e-12 ***
## season_Summer      27.874      6.450   4.322 1.56e-05 ***
## season_Fall        69.479      5.501  12.630  < 2e-16 ***
## year_2012          84.610      1.750  48.354  < 2e-16 ***
## month_Feb           4.343      4.413   0.984 0.325070
## month_Mar          15.021      4.975   3.019 0.002539 **
## month_Apr           7.179      7.331   0.979 0.327474
## month_May          18.652      7.846   2.377 0.017461 *
## month_Jun           4.688      8.073   0.581 0.561433
## month_Jul         -12.810      9.066  -1.413 0.157722
## month_Aug           5.286      8.818   0.600 0.548831
## month_Sep          30.154      7.850   3.842 0.000123 ***
## month_Oct           9.354      7.278   1.285 0.198725
## month_Nov         -10.060      7.018  -1.433 0.151761
## month_Dec          -7.715      5.599  -1.378 0.168201
## day_Monday        -20.511      9.447  -2.171 0.029925 *
## day_Tuesday       -20.389      9.898  -2.060 0.039417 *
## day_Wednesday     -17.930      9.893  -1.812 0.069943 .
## day_Thursday      -19.338      9.891  -1.955 0.050574 .
## day_Friday        -12.890      9.887  -1.304 0.192332
## day_Saturday       13.752      3.234   4.252 2.13e-05 ***
## `hour_1:00PM`     179.070      6.176  28.994  < 2e-16 ***
## `hour_10:00AM`    125.499      6.065  20.691  < 2e-16 ***
## `hour_10:00PM`     86.468      5.998  14.417  < 2e-16 ***
## `hour_11:00AM`    148.821      6.122  24.309  < 2e-16 ***
## `hour_11:00PM`     47.932      5.999   7.991 1.45e-15 ***
```

```
## `hour_12:00AM`     15.796      5.965   2.648 0.008101 **
## `hour_12:00PM`    191.865      6.171  31.091  < 2e-16 ***
## `hour_2:00AM`     -11.014      6.006  -1.834 0.066709 .
## `hour_2:00PM`     168.063      6.256  26.864  < 2e-16 ***
## `hour_3:00AM`     -18.782      6.043  -3.108 0.001887 **
## `hour_3:00PM`     177.468      6.213  28.564  < 2e-16 ***
## `hour_4:00AM`     -26.298      6.086  -4.321 1.57e-05 ***
## `hour_4:00PM`     240.870      6.241  38.594  < 2e-16 ***
## `hour_5:00AM`      -7.228      6.026  -1.199 0.230395
## `hour_5:00PM`     398.972      6.238  63.960  < 2e-16 ***
## `hour_6:00AM`      53.210      6.011   8.852  < 2e-16 ***
## `hour_6:00PM`     356.790      6.159  57.933  < 2e-16 ***
## `hour_7:00AM`     183.820      6.026  30.505  < 2e-16 ***
## `hour_7:00PM`     252.485      6.087  41.477  < 2e-16 ***
## `hour_8:00AM`     327.911      5.994  54.704  < 2e-16 ***
## `hour_8:00PM`     172.494      6.053  28.499  < 2e-16 ***
## `hour_9:00AM`     179.511      6.042  29.709  < 2e-16 ***
## `hour_9:00PM`     121.847      6.042  20.168  < 2e-16 ***
## `weather_Type 2`  -10.699      2.150  -4.977 6.54e-07 ***
## `weather_Type 3`  -65.089      3.630 -17.930  < 2e-16 ***
## `weather_Type 4`  -62.538     58.967  -1.061 0.288905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.8 on 13833 degrees of freedom
## Multiple R-squared:  0.6862, Adjusted R-squared:  0.6851
## F-statistic: 581.8 on 52 and 13833 DF,  p-value: < 2.2e-16
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  full_model$residuals
## D = 0.51877, p-value < 2.2e-16
## alternative hypothesis: two-sided

## [1] "H0 rejected: Residuals are NOT normally distributed"
```

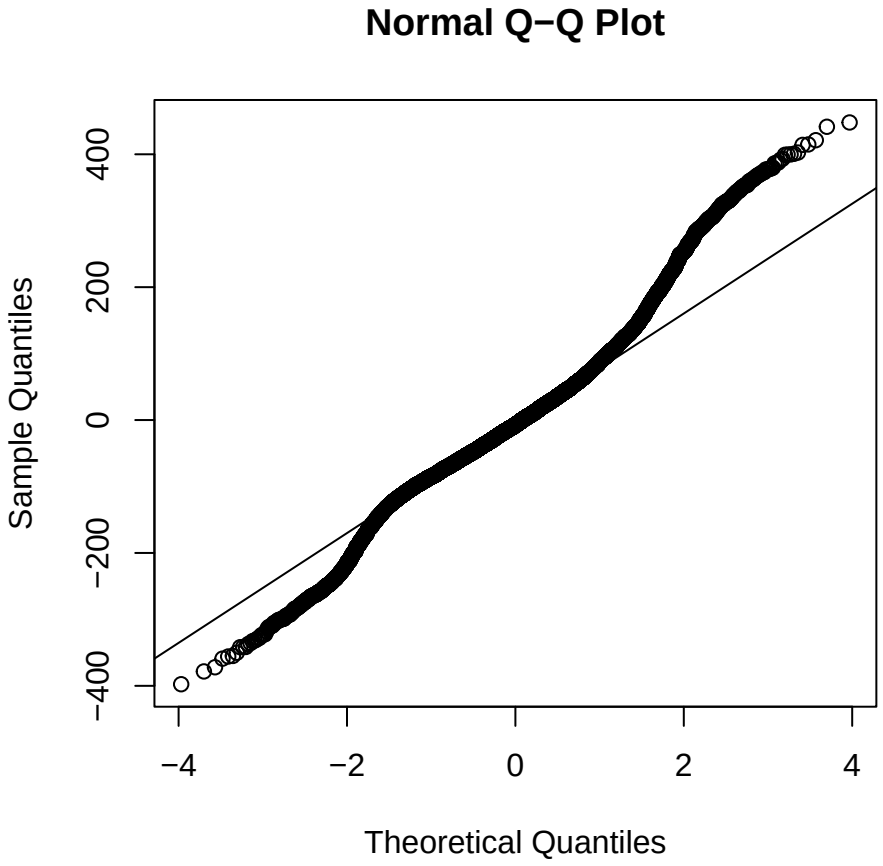

Figure 16: Normality of Residuals Assumption - Violated

```
## [1] "H0 rejected: Error variance spread INCONSTANTLY (Heteroscedasticity)"
```

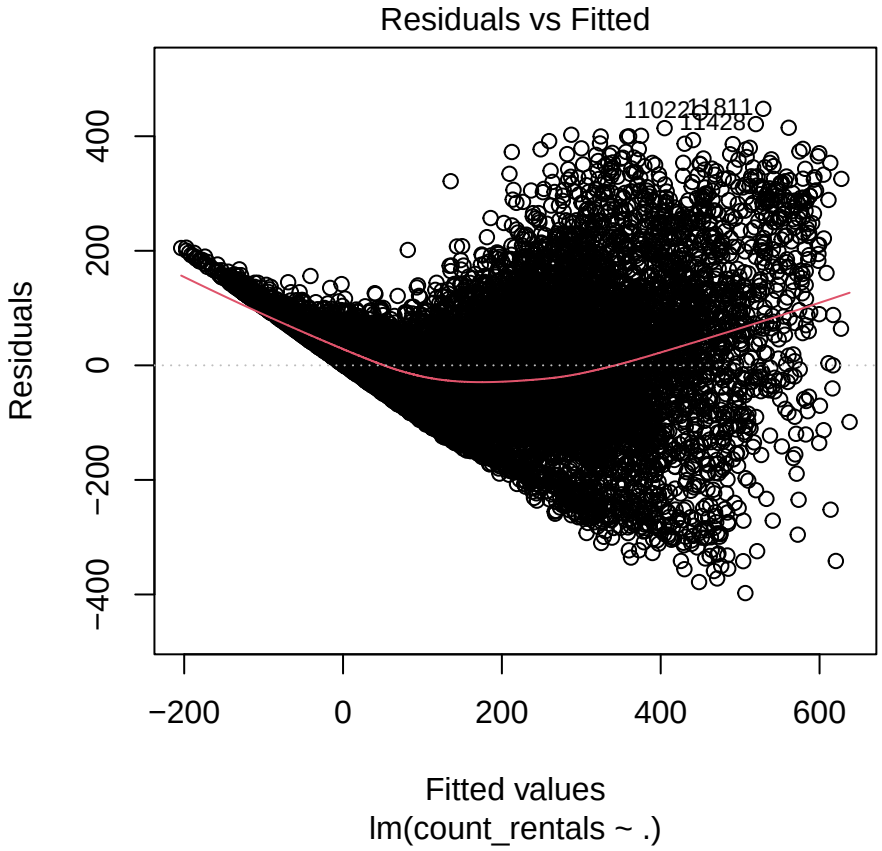

Figure 17: Homoscedasticity/Linearity/Independence of Residuals Assumptions - Violated

```
## [1] Below: Variance Inflation Factors of Full Model
```

	VIF
workingday	26.739507
day_Wednesday	16.181972
day_Tuesday	16.130844
day_Thursday	15.897965
day_Friday	15.689228
day_Monday	14.797938
season_Summer	10.626903
month_Jul	8.538623
month_Aug	8.001334
season_Spring	7.542580
season_Fall	7.514512
month_Jun	6.617642
month_May	6.542562
month_Sep	6.226942
month_Apr	5.552461
month_Oct	5.549132
temp	5.437115
month_Nov	5.041664
holiday	3.479981
month_Dec	3.240746
month_Mar	2.569427
`hour_3:00PM`	2.131679
`hour_4:00PM`	2.099477
`hour_1:00PM`	2.096349
`hour_2:00PM`	2.088711
`hour_5:00PM`	2.055942
`hour_6:00PM`	2.040987
`hour_12:00PM`	2.032408
`hour_7:00PM`	2.003863
`hour_11:00AM`	1.987010
`hour_8:00PM`	1.974601
`hour_12:00AM`	1.955377
`hour_10:00AM`	1.953571
`hour_10:00PM`	1.951465
`hour_11:00PM`	1.942585
`hour_9:00PM`	1.941474
`hour_8:00AM`	1.939842
`hour_6:00AM`	1.937930
`hour_2:00AM`	1.928431
`hour_5:00AM`	1.928420
`hour_9:00AM`	1.925865
`hour_7:00AM`	1.925018

```
## hum               1.922595
## `hour_3:00AM`     1.909999
## `hour_4:00AM`     1.891634
## month_Feb         1.854343
## day_Saturday      1.724501
## `weather_Type 3`  1.310499
## `weather_Type 2`  1.190148
## windspeed         1.179419
## year_2012         1.025737
## `weather_Type 4`  1.006424
```

```
## [1] ---
```

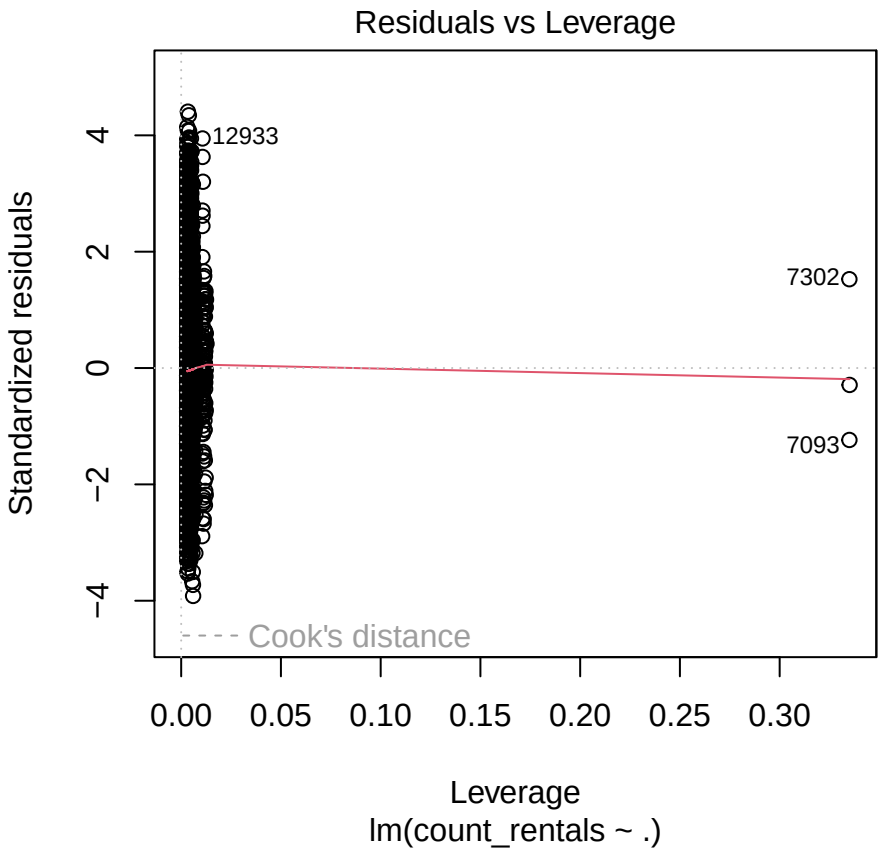

Figure 18: High Influential Observations

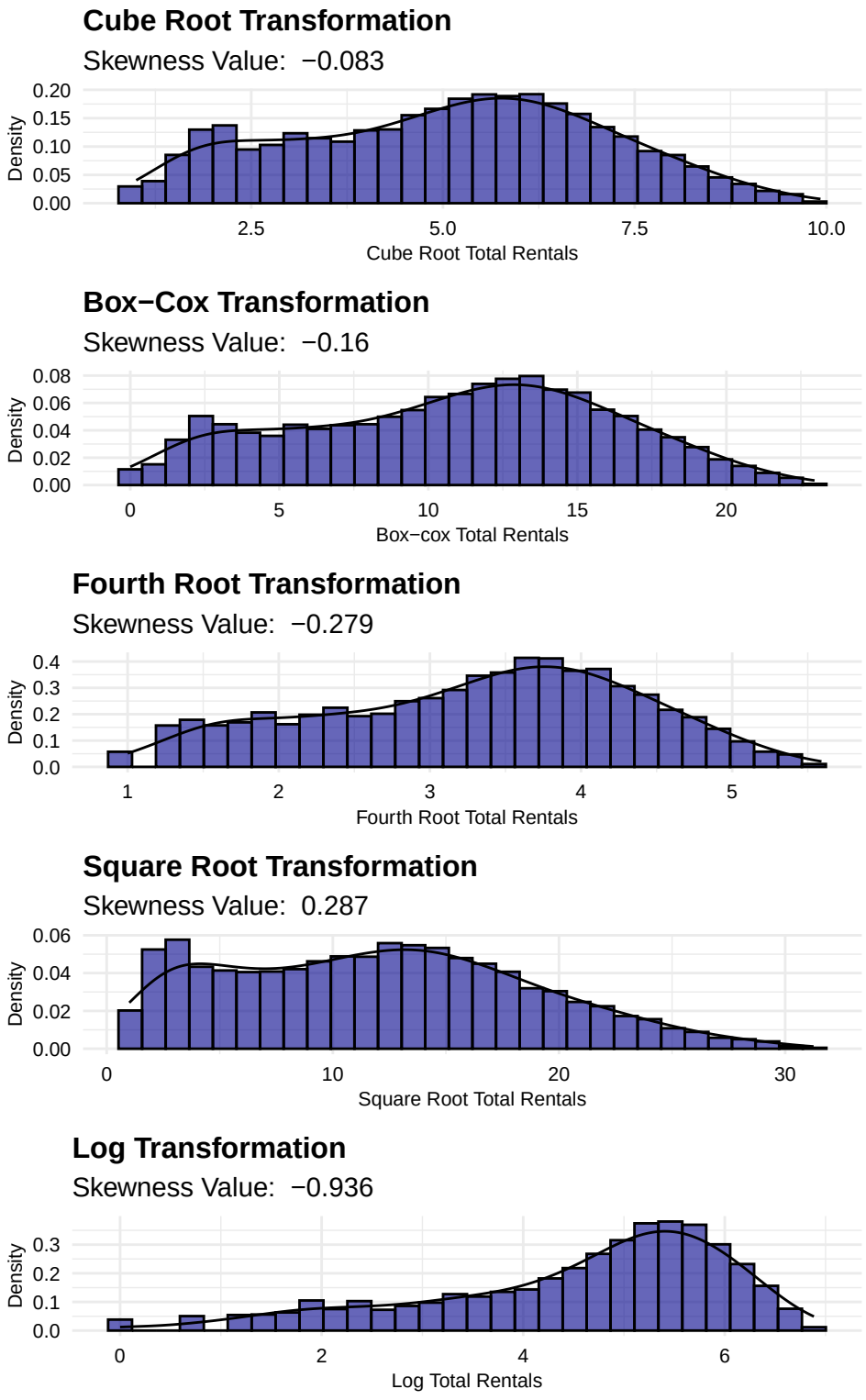

Figure 19: Cube Root Transformation Closest Skewness Value to Zero

Final Model Summary with $\hat{\beta}$ Coefficient Estimates

```
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = cube_root_total ~ ., data = train_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6841 -0.5236 -0.0027  0.6080  2.7782
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.43461    0.05851  24.519  < 2e-16 ***
## holiday           -0.31181    0.04317  -7.223 5.34e-13 ***
## temp               2.92401    0.04742  61.660  < 2e-16 ***
## hum               -0.63061    0.05373 -11.737  < 2e-16 ***
## windspeed         -0.38026    0.06659  -5.710 1.15e-08 ***
## season_Spring      0.32947    0.02251  14.633  < 2e-16 ***
## season_Fall        0.63961    0.01969  32.487  < 2e-16 ***
## year_2012          0.77631    0.01521  51.024  < 2e-16 ***
## month_May          0.12086    0.03205   3.771 0.000164 ***
## month_Aug          0.15035    0.03161   4.756 1.99e-06 ***
## month_Sep          0.32830    0.02993  10.967  < 2e-16 ***
## day_Friday         0.16198    0.02215   7.313 2.75e-13 ***
## day_Saturday       0.11866    0.02189   5.421 6.04e-08 ***
## `hour_1:00PM`      2.79024    0.05347  52.179  < 2e-16 ***
## `hour_10:00AM`     2.32341    0.05292  43.905  < 2e-16 ***
## `hour_10:00PM`     1.88597    0.05236  36.017  < 2e-16 ***
## `hour_11:00AM`     2.52961    0.05329  47.467  < 2e-16 ***
## `hour_11:00PM`     1.31734    0.05240  25.141  < 2e-16 ***
## `hour_12:00AM`     0.61021    0.05211  11.709  < 2e-16 ***
## `hour_12:00PM`     2.90408    0.05360  54.182  < 2e-16 ***
## `hour_2:00AM`     -0.40571    0.05248  -7.730 1.15e-14 ***
## `hour_2:00PM`      2.66734    0.05407  49.335  < 2e-16 ***
## `hour_3:00AM`     -0.85745    0.05280 -16.241  < 2e-16 ***
## `hour_3:00PM`      2.76112    0.05364  51.479  < 2e-16 ***
## `hour_4:00AM`     -1.10941    0.05315 -20.871  < 2e-16 ***
## `hour_4:00PM`      3.28997    0.05393  61.000  < 2e-16 ***
## `hour_5:00AM`     -0.26637    0.05262  -5.062 4.20e-07 ***
## `hour_5:00PM`      4.23614    0.05397  78.494  < 2e-16 ***
## `hour_6:00AM`      1.12898    0.05248  21.511  < 2e-16 ***
## `hour_6:00PM`      4.00441    0.05346  74.906  < 2e-16 ***
## `hour_7:00AM`      2.59690    0.05263  49.342  < 2e-16 ***
```

```
## `hour_7:00PM`     3.37096    0.05293  63.688  < 2e-16 ***
## `hour_8:00AM`     3.77600    0.05238  72.085  < 2e-16 ***
## `hour_8:00PM`     2.77559    0.05274  52.624  < 2e-16 ***
## `hour_9:00AM`     2.92134    0.05278  55.346  < 2e-16 ***
## `hour_9:00PM`     2.28536    0.05270  43.368  < 2e-16 ***
## `weather_Type 2` -0.08979    0.01871  -4.799 1.61e-06 ***
## `weather_Type 3` -0.80743    0.03162 -25.538  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8893 on 13845 degrees of freedom
## Multiple R-squared:  0.8088, Adjusted R-squared:  0.8083
## F-statistic:  1583 on 37 and 13845 DF,  p-value: < 2.2e-16
```

Final Model Equation Predicting Cube Root of Rentals

Corresponding $\hat{\beta}$ Coefficients Above(Model Summary)

$$\hat{cuberoottotal} = \hat{\beta}_0 + \hat{\beta}_1\texttt{holiday} + \hat{\beta}_2\texttt{temp} + \hat{\beta}_3\texttt{hum} + \hat{\beta}_4\texttt{season\_Spring} + \hat{\beta}_5\texttt{season\_Fall} + \hat{\beta}_6\texttt{year\_2012} + \hat{\beta}_7\texttt{month\_May} + \hat{\beta}_8\texttt{month\_Aug} + \hat{\beta}_9\texttt{month\_Sep} + \hat{\beta}_{10}\texttt{day\_Friday} + \hat{\beta}_{11}\texttt{day\_Saturday} + \hat{\beta}_{12}\texttt{hour\_1:00PM} + \hat{\beta}_{13}\texttt{hour\_10:00AM} + \hat{\beta}_{14}\texttt{hour\_10:00PM} + \hat{\beta}_{15}\texttt{hour\_11:00AM} + \hat{\beta}_{16}\texttt{hour\_11:00PM} + \hat{\beta}_{17}\texttt{hour\_12:00AM} + \hat{\beta}_{18}\texttt{hour\_12:00PM} + \hat{\beta}_{19}\texttt{hour\_2:00AM} + \hat{\beta}_{20}\texttt{hour\_2:00PM} + \hat{\beta}_{21}\texttt{hour\_3:00AM} + \hat{\beta}_{22}\texttt{hour\_3:00PM} + \hat{\beta}_{23}\texttt{hour\_4:00AM} + \hat{\beta}_{24}\texttt{hour\_4:00PM} + \hat{\beta}_{25}\texttt{hour\_5:00AM} + \hat{\beta}_{26}\texttt{hour\_5:00PM} + \hat{\beta}_{27}\texttt{hour\_6:00AM} + \hat{\beta}_{28}\texttt{hour\_6:00PM} + \hat{\beta}_{29}\texttt{hour\_7:00AM} + \hat{\beta}_{30}\texttt{hour\_7:00PM} + \hat{\beta}_{31}\texttt{hour\_8:00AM} + \hat{\beta}_{32}\texttt{hour\_8:00PM} + \hat{\beta}_{33}\texttt{hour\_9:00AM} + \hat{\beta}_{34}\texttt{hour\_9:00PM} + \hat{\beta}_{35}\texttt{weather\_Type\_2} + \hat{\beta}_{36}\texttt{weather\_Type\_3} + \epsilon$$

```
## [1] Below: Variance Inflation Factors of Final Model
```

```
##          holiday             temp              hum        windspeed
##         1.010373         1.450399         1.884040         1.160856
##    season_Spring      season_Fall        year_2012        month_May
##         1.696245         1.260991         1.015778         1.430472
##        month_Aug        month_Sep       day_Friday     day_Saturday
##         1.347341         1.186464         1.031496         1.034603
##    `hour_1:00PM`   `hour_10:00AM`   `hour_10:00PM`   `hour_11:00AM`
##         2.058933         1.948351         1.948948         1.972523
##   `hour_11:00PM`   `hour_12:00AM`   `hour_12:00PM`    `hour_2:00AM`
##         1.941928         1.955579         2.008643         1.929251
##    `hour_2:00PM`    `hour_3:00AM`    `hour_3:00PM`    `hour_4:00AM`
##         2.043812         1.910181         2.081389         1.890195
##    `hour_4:00PM`    `hour_5:00AM`    `hour_5:00PM`    `hour_6:00AM`
##         2.050717         1.926402         2.016169         1.935479
##    `hour_6:00PM`    `hour_7:00AM`    `hour_7:00PM`    `hour_8:00AM`
##         2.011446         1.923911         1.984768         1.940809
##    `hour_8:00PM`    `hour_9:00AM`    `hour_9:00PM` `weather_Type 2`
##         1.964450         1.925360         1.935191         1.180836
## `weather_Type 3`
##         1.302384
```

```
## [1] ---
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  stepwise_model$residuals
## D = 0.05551, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  studentized Breusch-Pagan test
##
## data:  stepwise_model
## BP = 3826, df = 37, p-value < 2.2e-16
```

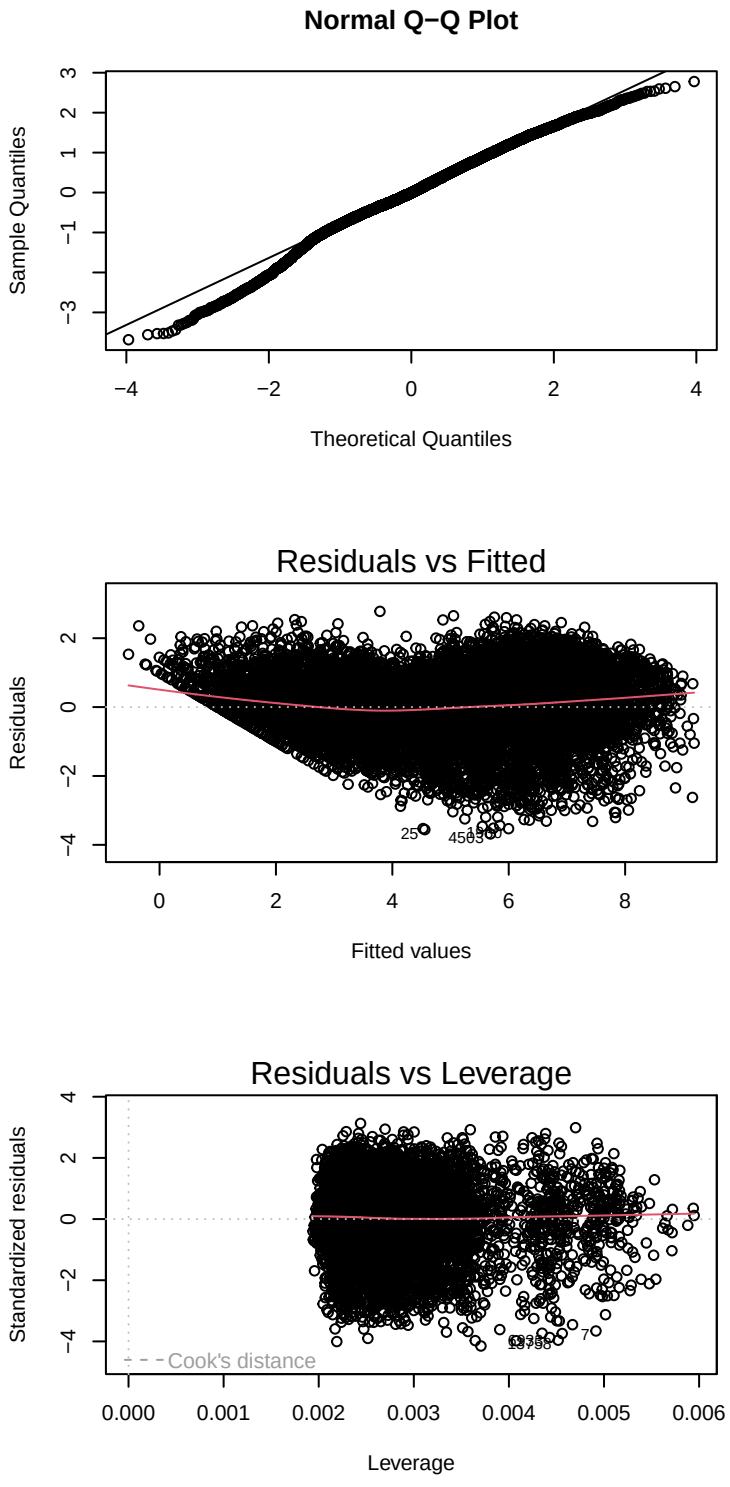

Figure 20: Assumption Corrections and High Leverage Removal

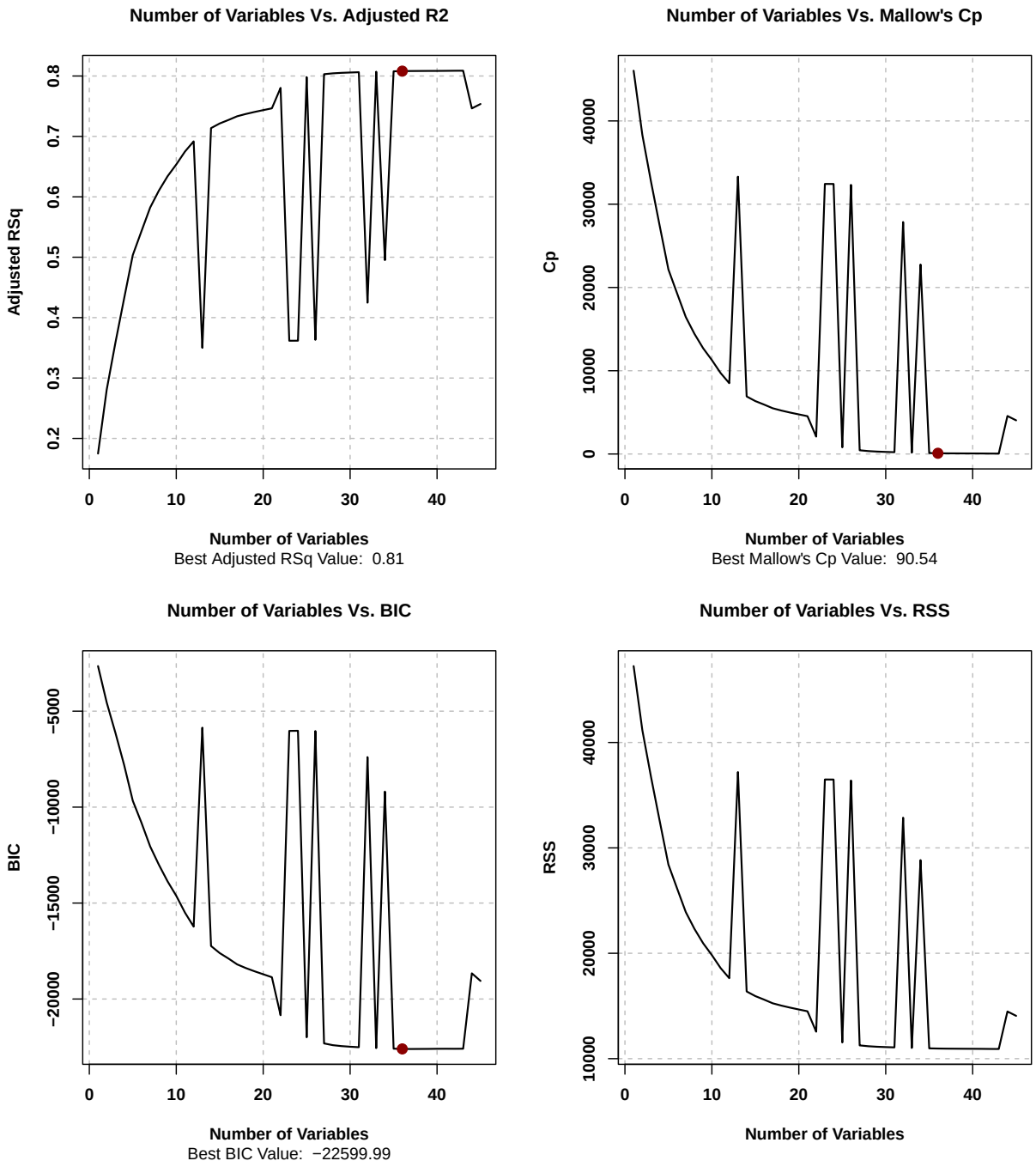

Figure 21: Stepwise Selection Model Evaluations(BIC/RSquared/Mallows Cp/RSS)

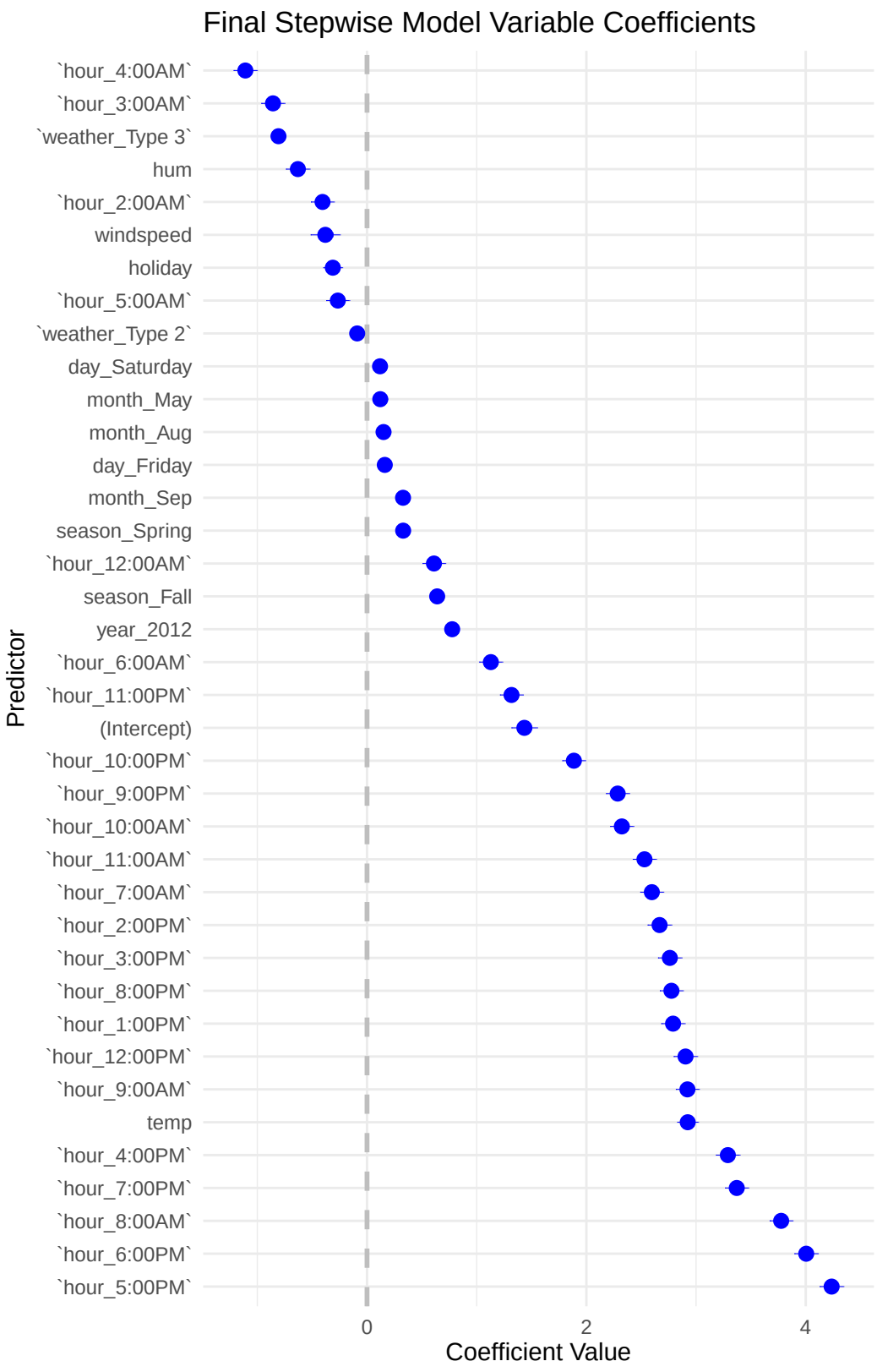

Figure 22: Predictors for Final Stepwise Model

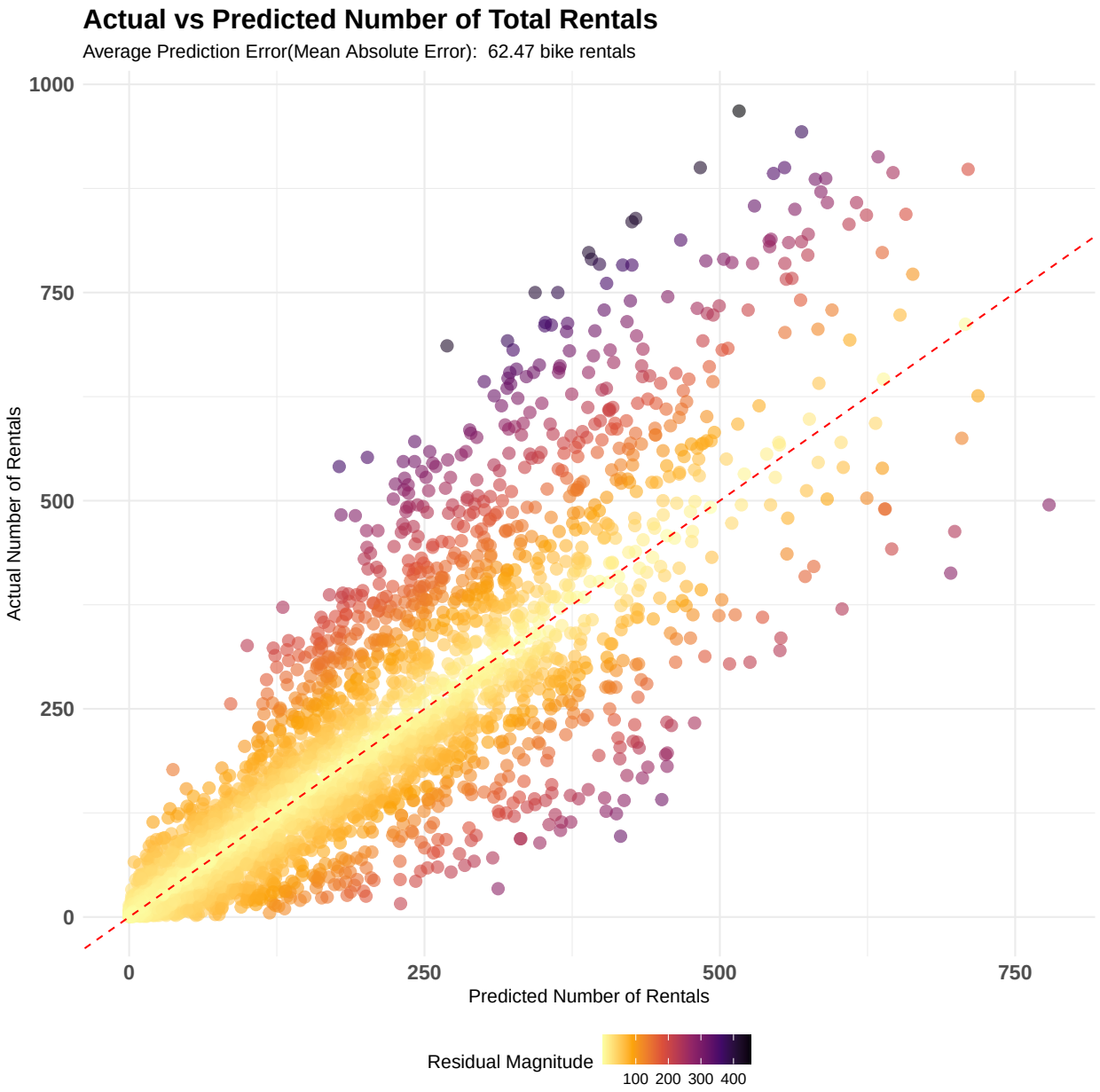

Figure 23: Predictions Using the Test Dataset